bio-protocol Preprint

Not Peer Reviewed
Ask a question online

Unsupervised clustering of 4,000-gene human MERFISH data

Rongxin Fang Xiaowei Zhuang

Updated date: Sep 8, 2022

An abbreviated version of this protocol was published in Science in Jun 2022

Conservation and divergence of cortical cell organization in human and mouse revealed by MERFISH

DOI: 10.1126/science.abm1741

Detailed protocol

Unsupervised clustering of 4,000-gene human MERFISH data

We used the RNA molecules identified in the nucleus for clustering analysis. With the nucleus-by-gene matrix, we first preprocessed the matrix by several steps. The segmentation approach generated a small fraction of putative “nuclei” with very small total volumes or low RNA abundance due to segmentation artifacts, as well as some cells that overlapped in the 3-D dimension and were not properly separated. We first filtered any cells with segmentation appeared in less than three z-planes to avoid spurious segmentations. Second, we filtered cells whose centroid was within 100 pixels to the edge of the FOV to avoid edge effect. Third, we removed the segmented “cells” that had a volume that was either less than 300um^2 or larger than 6000 um^2. Fourth, we calculated the RNA molecule density in each nucleus and filtered out instances with density less than 0.1 molecule / um^2, which was closed to the molecule density outside the cell body, likely representing “empty nucleus” introduced by spurious segmentations or out-of-focus signal.

Using the remaining nuclei, we next performed clustering analysis. In detail, we first normalized the nucleus-by-gene count matrix using “scTransform”, a modeling framework for the normalization and variance stabilization of molecular count data from scRNA-seq experiments. To remove the differences in RNA counts due to the incompleteness of nucleus, we further normalized the RNA counts per cell by regressing out the imaged volume of each cell using “scTransform” by setting the “vars.to.regress” to the area size of the nucleus. Next, we performed dimensionality reduction using principal component analysis (PCA), restricted to the 30 principal components (PCs) with the highest eigenvalues, and finally visualized using a 2D UMAP embedding. In the UMAP space, the batch effect between MERFISH experiments was further eliminated using Canonical Correlation Analysis (CCA) which was implemented in Seurat V3. To identify transcriptionally distinct cell clusters, we performed graph-based Leiden community detection (k=15; resolution=0.5) in the 30 PCs-space, unless otherwise mentioned.

We first annotated identified clusters to major cell types based on the expression of canonical marker genes. Next, we performed separate clustering of inhibitory neurons and excitatory neurons from MERFISH data and SMART-seq data (https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq) independently using the Single-Cell Consensus Clustering method (SC3) to examine the correspondence between the two data modalities. We next determined the excitatory and inhibitory neuronal clusters by integrated analysis of the SMART-seq and MERFISH data. In detail, we performed normalization using “scTransform” for MERFISH and SMART-seq independently. ScTransform automatically selected 2,000 variable genes in both datasets as “anchors” for integration. Second, we used the selected anchors and CCA to generate the joint low-dimensional embedding space between MERFISH and SMART-seq data. In the joint space, we identified clusters using graph-based clustering method (Leiden). A key parameter in Leiden is resolution, a higher resolution usually resulting in more clusters. How to choose the optimal resolution is described with more details below. First, for a clustering result generated by a given resolution, we trained a KNN classifier in the joint embedding space on 80% of the dataset and then estimated the prediction accuracy on the remaining 20%. We conducted this five times (also known as five-fold cross validation) to estimate the averaged prediction accuracy. As expected, we observed a decrease in prediction accuracy with increasing resolution, suggesting “over splitting” of clusters. In this study, we chose the resolution that yield 3% misidentification rate. To avoid the errors introduced by the integration algorithm, we next filtered any “outlier” clusters that were mostly identified by one dataset. In detail, for each cluster, we calculated the ratio between cell proportions determined by SMART-seq and MERFISH and then normalized these ratios to z-score across all clusters. We removed any clusters whose z-score was larger than 3 or smaller than -3. Finally, for each cluster, we performed differential expression (DE) analysis using Wilcoxon Rank Sum test in MERFISH and SMART-seq data separately (P-value < 1e-2; fold-change > 1.2). A cluster which failed to find a DE gene in both MERFISH and SMART-seq was merged with the closest clusters in the joint embedding space. This was repeated until no clusters could be merged anymore. For clustering of non-neuronal cells, we used MERFISH data alone, because these cells were depleted in SMART-seq dataset

How to cite:(Readers should cite both the Bio-protocol preprint and the original research article where this protocol was used)

1. Fang, R. and Zhuang, X. (2022). Unsupervised clustering of 4,000-gene human MERFISH data. Bio-protocol Preprint. bio-protocol.org/prep1915.
2. Fang, R., Xia, C., Close, J. L., Zhang, M., He, J., Huang, Z., Halpern, A. R., Long, B., Miller, J. A., Lein, E. S. and Zhuang, X.(2022). Conservation and divergence of cortical cell organization in human and mouse revealed by MERFISH. Science 377(6601). DOI: 10.1126/science.abm1741